ivtobit — Tobit model with continuous endogenous covariates

Description

ivtobit fits tobit models where one or more of the covariates are endogenously determined. By default, ivtobit uses maximum likelihood estimation, but Newey’s (1987) minimum chi-squared (two-step) estimator can be requested. Both estimators assume that the endogenous covariates are continuous and so are not appropriate for use with discrete endogenous covariates.

Quick start

Tobit regression of \( y_1 \) on \( x \) and endogenous regressor \( y_2 \) that is instrumented by \( z \) where \( y_1 \) is left-censored at its observed minimum

\[
\text{ivtobit} \ y_1 \ x \ (y_2 = z), \ ll
\]

As above, but specify that \( y_1 \) is left-censored at 0 and right-censored at 20

\[
\text{ivtobit} \ y_1 \ x \ (y_2 = z), \ ll(0) \ ul(20)
\]

Use Newey’s two-step estimator

\[
\text{ivtobit} \ y_1 \ x \ (y_2 = z), \ ll(0) \ ul(20) \ twostep
\]

As above, and show first-stage regression results

\[
\text{ivtobit} \ y_1 \ x \ (y_2 = z), \ ll(0) \ ul(20) \ twostep \ first
\]

Menu

Statistics > Endogenous covariates > Tobit model with endogenous covariates
Syntax

**Maximum likelihood estimator**

```
ivtobit depvar [varlist1] (varlist2 = varlist_{iv}) [if] [in] [weight],
   ll[(#)] ul[(#)] [mle_options]
```

**Two-step estimator**

```
ivtobit depvar [varlist1] (varlist2 = varlist_{iv}) [if] [in] [weight], twostep
   ll[(#)] ul[(#)] [tse_options]
```

`varlist1` is the list of exogenous variables.

`varlist2` is the list of endogenous variables.

`varlist_{iv}` is the list of exogenous variables used with `varlist1` as instruments for `varlist2`.

### `mle_options`

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td><code>*ll[(#)]</code></td>
</tr>
<tr>
<td><code>*ul[(#)]</code></td>
</tr>
<tr>
<td><code>mle</code></td>
</tr>
<tr>
<td><code>constraints(constraints)</code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE/Robust</td>
</tr>
<tr>
<td><code>vce(vcetype)</code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reporting</td>
</tr>
<tr>
<td><code>level(#)</code></td>
</tr>
<tr>
<td><code>first</code></td>
</tr>
<tr>
<td><code>nocnsreport</code></td>
</tr>
<tr>
<td><code>display_options</code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximization</td>
</tr>
<tr>
<td><code>maximize_options</code></td>
</tr>
<tr>
<td><code>coeflegend</code></td>
</tr>
</tbody>
</table>

*You must specify at least one of `ll[(#)]` and `ul[(#)]`. |
### tse_options

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
</tr>
<tr>
<td><em>twostep</em></td>
</tr>
<tr>
<td>*ll [(#)]</td>
</tr>
<tr>
<td>*ul [(#)]</td>
</tr>
<tr>
<td><strong>SE</strong></td>
</tr>
<tr>
<td>vce(vcetype)</td>
</tr>
<tr>
<td><strong>Reporting</strong></td>
</tr>
<tr>
<td>level(#)</td>
</tr>
<tr>
<td>first</td>
</tr>
<tr>
<td>display_options</td>
</tr>
<tr>
<td>coeflegend</td>
</tr>
</tbody>
</table>

*twostep* is required. You must specify at least one of ll[(#)] and ul[(#)].

### Options for ML estimator

- **Model**
  - ll[(#)] and ul[(#)] indicate the lower and upper limits for censoring, respectively. You may specify one or both. Observations with \( \text{depvar} \leq \text{ll()} \) are left-censored; observations with \( \text{depvar} \geq \text{ul()} \) are right-censored; and remaining observations are not censored. You do not have to specify the censoring values at all. It is enough to type ll, ul, or both. When you do not specify a censoring value, ivtobit assumes that the lower limit is the minimum observed in the data (if ll is specified) and that the upper limit is the maximum (if ul is specified).
  - mle requests that the conditional maximum-likelihood estimator be used. This is the default.
- **SE/Robust**
  - vce(vcetype) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (oim, opg), that are robust to some kinds of misspecification (robust), that allow for intragroup correlation (cluster clustvar), and that use bootstrap or jackknife methods (bootstrap, jackknife); see [R] vce_option.
**ivtobit — Tobit model with continuous endogenous covariates**

**Reporting**

level(#) ; see [R] Estimation options.

*first* requests that the parameters for the reduced-form equations showing the relationships between the endogenous variables and instruments be displayed. For the two-step estimator, *first* shows the first-stage regressions. For the maximum likelihood estimator, these parameters are estimated jointly with the parameters of the tobit equation. The default is not to show these parameter estimates.

*nocnsreport* ; see [R] Estimation options.

display_options: noci, nopvalues, noomitted, vsquish, noemptycells, baselevels, allbaselevels, nolabel, fwrap(#), fwrapon(style), cformat(%fmt), pformat(%fmt), sformat(%fmt), and nolstretch; see [R] Estimation options.

**Maximization**

maximize_options: difficult, technique(algorithm_spec), iterate(#), [no]log, trace, gradient, showstep, hessian, showtolerance, tolerance(#), ltolerance(#), nrtolerance(#), nonrtolerance, and from(init_specs); see [R] Maximize.

Setting the optimization type to technique(bhhh) resets the default vcetype to vce(opg).

The following option is available with ivtobit but is not shown in the dialog box:

*coeflegend* ; see [R] Estimation options.

**Options for two-step estimator**

**Model**

twostep is required and requests that Newey’s (1987) efficient two-step estimator be used to obtain the coefficient estimates.

11[(#) ] and u1[(#) ] indicate the lower and upper limits for censoring, respectively. You may specify one or both. Observations with *depvar* ≤ 11() are left-censored; observations with *depvar* ≥ u1() are right-censored; and remaining observations are not censored. You do not have to specify the censoring values at all. It is enough to type 11, u1, or both. When you do not specify a censoring value, ivtobit assumes that the lower limit is the minimum observed in the data (if 11 is specified) and that the upper limit is the maximum (if u1 is specified).

**SE**

vce(vcetype) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (twostep) and that use bootstrap or jackknife methods (bootstrap, jackknife); see [R] vce_option.

**Reporting**

level(#) ; see [R] Estimation options.

*first* requests that the parameters for the reduced-form equations showing the relationships between the endogenous variables and instruments be displayed. For the two-step estimator, *first* shows the first-stage regressions. For the maximum likelihood estimator, these parameters are estimated jointly with the parameters of the tobit equation. The default is not to show these parameter estimates.
**vtobit** — Tobit model with continuous endogenous covariates

- `display_options`: `noci`, `nopvalues`, `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(#)`, `fvwrapon(style)`, `cformat(%fmt)`, `pformat(%fmt)`, `sformat(%fmt)`, and `nolstretch`; see [R] Estimation options.

The following option is available with `ivtobit` but is not shown in the dialog box:

`coeflegend`; see [R] Estimation options.

### Remarks and examples

`ivtobit` fits models with censored dependent variables and endogenous covariates. You can use it to fit a tobit model when you suspect that one or more of the covariates is correlated with the error term. `ivtobit` is to tobit what `ivregress` is to linear regression analysis; see [R] `ivregress` for more information.

Formally, the model is

\[
y_{1i}^* = y_{2i}\beta + x_{1i}\gamma + u_i \\
y_{2i} = x_{1i}\Pi_1 + x_{2i}\Pi_2 + v_i
\]

where \(i = 1, \ldots, N\); \(y_{2i}\) is a \(1 \times p\) vector of endogenous variables; \(x_{1i}\) is a \(1 \times k_1\) vector of exogenous variables; \(x_{2i}\) is a \(1 \times k_2\) vector of additional instruments; and the equation for \(y_{2i}\) is written in reduced form. By assumption \((u_i, v_i) \sim N(0, \Sigma)\). \(\beta\) and \(\gamma\) are vectors of structural parameters, and \(\Pi_1\) and \(\Pi_2\) are matrices of reduced-form parameters. We do not observe \(y_{1i}^*\); instead, we observe

\[
y_{1i} = \begin{cases} 
  a & y_{1i}^* < a \\
  y_{1i}^* & a \leq y_{1i}^* \leq b \\
  b & y_{1i}^* > b
\end{cases}
\]

The order condition for identification of the structural parameters is that \(k_2 \geq p\). Presumably, \(\Sigma\) is not block diagonal between \(u_i\) and \(v_i\); otherwise, \(y_{2i}\) would not be endogenous.

#### Technical note

This model is derived under the assumption that \((u_i, v_i)\) is independent and identically distributed multivariate normal for all \(i\). The `vce(cluster clustvar)` option can be used to control for a lack of independence. As with the standard tobit model without endogeneity, if \(u_i\) is heteroskedastic, point estimates will be inconsistent.

#### Example 1: Estimation and parameter interpretation

We model the number of hours per week that high school boys spend using social media (hsocial). The data collection process caused the observations on the number of hours spent to be censored at 12 hours. A tobit-type model is therefore reasonable for our data.

We model each boy’s number of hours spent using social media as a function of whether he has a smartphone (sphone), whether he has a computer at home (computer), the year in high school in which he is enrolled (year), and the hours per week he spends studying (hstudy).

We believe that there are unobservable variables that simultaneously affect hstudy and hsocial, which is to say that hstudy is endogenous. Because hstudy is endogenous, we must model it as well. Our model for the endogenous hstudy always includes the exogenous covariates used to model the outcome hsocial. We must also include at least one covariate in the model for the endogenous hstudy that was not included in the model for the outcome hsocial.
We use ivtobit with the default maximum-likelihood estimator to model the endogenous variable `hstudy` as a function of the highest educational degree attained by their parents (`pedu`), the time spent watching television (`tvhours`), and the exogenous covariates used to model `hsocial`.

```stata
use https://www.stata-press.com/data/r16/smedia
ivtobit hsocial i.sphone i.computer i.year (hstudy = tvhours i.pedu), ul(12)
```

Fitting exogenous tobit model

Fitting full model

```
Iteration 0: log likelihood = -3240.5279
Iteration 1: log likelihood = -3186.8824
Iteration 2: log likelihood = -3173.1147
Iteration 3: log likelihood = -3172.8561
Iteration 4: log likelihood = -3172.856
Tobit model with endogenous regressors
Number of obs = 1,324
Uncensored = 928
Limits: lower = -inf Left-censored = 0
upper = 12 Right-censored = 396
Wald chi2(6) = 11610.73
Log likelihood = -3172.856 Prob > chi2 = 0.0000
```

The coefficients in the table tell us how much the linear prediction for the outcome changes when there is a change in a covariate. Unlike the tobit model, where the linear prediction is the expected value of the outcome as if the data had not been censored, we need to incorporate the other parameters that were estimated by ivtobit to obtain effects that account for endogeneity.

We recommend that you use `margins` to estimate the effect of a covariate on the mean of the outcome given the covariates. Say we want to estimate the effect of all boys having a smartphone relative to the case where no boy has a smartphone on `hsocial`.

```
The coefficients in the table tell us how much the linear prediction for the outcome changes when there is a change in a covariate. Unlike the tobit model, where the linear prediction is the expected value of the outcome as if the data had not been censored, we need to incorporate the other parameters that were estimated by ivtobit to obtain effects that account for endogeneity.

We recommend that you use `margins` to estimate the effect of a covariate on the mean of the outcome given the covariates. Say we want to estimate the effect of all boys having a smartphone relative to the case where no boy has a smartphone on `hsocial`.
```
. margins, dydx(sphone) predict(ystar(.,.))

Average marginal effects
Number of obs = 1,324
Model VCE : OIM
Expression : E(hsocial), predict(ystar(.,.))
dy/dx w.r.t. : 1.sphone

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dy/dx</td>
</tr>
<tr>
<td>1.sphone</td>
<td>6.254632</td>
</tr>
</tbody>
</table>

Note: dy/dx for factor levels is the discrete change from the base level.

The effect is 6.3 more hours a week of social media usage. This is different from the coefficient in the table, which has a value of 6. See [R] ivtobit postestimation for additional examples.

Below the table, we see a Wald test for whether the correlation between the residuals from the main equation (predicting hstudy) and the residuals from the auxiliary equation (predicting hsocial) is 0. The correlation itself is 0.47 and shown in the table as corr(e.hstudy,e.hsocial). If the test statistic is not significant, there is not sufficient information in the sample to reject the null hypothesis of no endogeneity. In our example, we reject the null hypothesis that supports our choice of a tobit model that accounts for endogeneity.

Technical note

In the tobit model with endogenous covariates, we assume that \((u_i, v_i)\) is multivariate normal with covariance matrix

\[
\text{Var}(u_i, v_i) = \Sigma = \begin{bmatrix}
\sigma_u^2 & \Sigma_{21} \\
\Sigma_{21} & \Sigma_{22}
\end{bmatrix}
\]

Using the properties of the multivariate normal distribution, \(\text{Var}(u_i|v_i) \equiv \sigma_{u|v}^2 = \sigma_u^2 - \Sigma_{21}^{-1} \Sigma_{22} \Sigma_{21}\). Calculating the marginal effects on the conditional expected values of the observed and latent dependent variables and on the probability of censoring requires an estimate of \(\sigma_{u|v}^2\). Unlike the default maximum-likelihood estimator, the two-step estimator identifies only \(\sigma_{u|v}^2\), not \(\sigma_u^2\), so only the linear prediction and its standard error are available after you have used the twostep option.
ivtobit, mle stores the following in e():

Scalars
\[ e(N) \] number of observations
\[ e(N\_unc) \] number of uncensored observations
\[ e(N\_lc) \] number of left-censored observations
\[ e(N\_rc) \] number of right-censored observations
\[ e(llopt) \] minimum of depvar or contents of ll()
\[ e(ulopt) \] minimum of depvar or contents of ul()
\[ e(k) \] number of parameters
\[ e(k\_eq) \] number of equations in e(b)
\[ e(k\_eq\_model) \] number of equations in overall model test
\[ e(k\_dv) \] number of dependent variables
\[ e(df\_m) \] model degrees of freedom
\[ e(ll) \] log likelihood
\[ e(N\_clust) \] number of clusters
\[ e(endog\_ct) \] number of endogenous covariates
\[ e(p) \] model Wald \( p \)-value
\[ e(p\_exog) \] exogeneity test Wald \( p \)-value
\[ e(chi2) \] model Wald \( \chi^2 \)
\[ e(chi2\_exog) \] Wald \( \chi^2 \) test of exogeneity
\[ e(rank) \] rank of e(V)
\[ e(ic) \] number of iterations
\[ e(rc) \] return code
\[ e(converged) \] 1 if converged, 0 otherwise

Macros
\[ e(cmd) \] ivtobit
\[ e(cmdline) \] command as typed
\[ e(depvar) \] name of dependent variable
\[ e(instd) \] instrumented variables
\[ e(insts) \] instruments
\[ e(wtype) \] weight type
\[ e(wexp) \] weight expression
\[ e(title) \] title in estimation output
\[ e(clustvar) \] name of cluster variable
\[ e(chi2type) \] Wald; type of model \( \chi^2 \) test
\[ e(vce) \] vcetype specified in vce()
\[ e(vcetype) \] title used to label Std. Err.
\[ e(method) \] ml
\[ e(opt) \] type of optimization
\[ e(which) \] max or min; whether optimizer is to perform maximization or minimization
\[ e(ml\_method) \] type of ml method
\[ e(user) \] name of likelihood-evaluator program
\[ e(technique) \] maximization technique
\[ e(properties) \] b V
\[ e(estat\_cmd) \] program used to implement estat
\[ e(predict) \] program used to implement predict
\[ e(footnote) \] program used to implement the footnote display
\[ e(marginsok) \] predictions allowed by margins
\[ e(marginsprop) \] signals to the margins command
\[ e(asbalanced) \] factor variables fvset as asbalanced
\[ e(asobserved) \] factor variables fvset as asobserved
Matrices
- **e(b)**: coefficient vector
- **e(Cns)**: constraints matrix
- **e(ilog)**: iteration log (up to 20 iterations)
- **e(gradient)**: gradient vector
- **e(Sigma)**: $\hat{\Sigma}$
- **e(V)**: variance–covariance matrix of the estimators
- **e(V_modelbased)**: model-based variance

Functions
- **e(sample)**: marks estimation sample

**ivtobit, twostep** stores the following in **e()**: 

Scalars
- **e(N)**: number of observations
- **e(N_unc)**: number of uncensored observations
- **e(N_lc)**: number of left-censored observations
- **e(N_rc)**: number of right-censored observations
- **e(llopt)**: contents of ll()
- **e(ulopt)**: contents of ul()
- **e(df_m)**: model degrees of freedom
- **e(df_exog)**: degrees of freedom for $\chi^2$ test of exogeneity
- **e(p)**: model Wald $p$-value
- **e(p_exog)**: exogeneity test Wald $p$-value
- **e(chi2)**: model Wald $\chi^2$
- **e(chi2_exog)**: Wald $\chi^2$ test of exogeneity
- **e(rank)**: rank of $e(V)$

Macros
- **e(cmd)**: ivtobit
- **e(cmdline)**: command as typed
- **e(depvar)**: name of dependent variable
- **e(instd)**: instrumented variables
- **e(insts)**: instruments
- **e(wtype)**: weight type
- **e(wexp)**: weight expression
- **e(chi2type)**: Wald; type of model $\chi^2$ test
- **e(vce)**: vcetype specified in vce()
- **e(method)**: twostep
- **e(properties)**: b V
- **e(estat_cmd)**: program used to implement estat
- **e(predict)**: program used to implement predict
- **e(footnote)**: program used to implement the footnote display
- **e(marginsok)**: predictions allowed by margins
- **e(marginsprop)**: signals to the margins command
- **e(asbalanced)**: factor variables fvset as asbalanced
- **e(asobserved)**: factor variables fvset as asobserved

Matrices
- **e(b)**: coefficient vector
- **e(Cns)**: constraints matrix
- **e(V)**: variance–covariance matrix of the estimators
- **e(V_modelbased)**: model-based variance

Functions
- **e(sample)**: marks estimation sample
Methods and formulas

The estimation procedure used by \texttt{ivtobit} is similar to that used by \texttt{ivprobit}. For compactness, we write the model as

\[ y^*_1 = z_i \delta + u_i \]

\[ y_{2i} = x_i \Pi + v_i \]

where \( z_i = (y_{2i}, x_i) \), \( x_i = (x_{1i}, x_{2i}) \), \( \delta = (\beta', \gamma')' \), and \( \Pi = (\Pi_1', \Pi_2')' \). We do not observe \( y^*_1 \); instead, we observe \( y_{1i} \)

\[ y_{1i} = \begin{cases} 
  a & y^*_1 < a \\
  y^*_1 & a \leq y^*_1 \leq b \\
  b & y^*_1 > b 
\end{cases} \]

(\( u_i, v_i \)) is distributed multivariate normal with mean zero and covariance matrix

\[ \Sigma = \begin{bmatrix} \sigma_u^2 & \Sigma_{u1} \\ \Sigma_{1u} & \Sigma_{11} \end{bmatrix} \]

Using the properties of the multivariate normal distribution, we can write \( u_i = v_i' \alpha + \epsilon_i \), where \( \alpha = \Sigma_{22}^{-1} \Sigma_{12} ; \epsilon_i \sim \mathcal{N}(0; \sigma_{u|v}^2) \), where \( \sigma_{u|v}^2 = \sigma_u^2 - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12} \); and \( \epsilon_i \) is independent of \( v_i, z_i, \) and \( x_i \).

The likelihood function is straightforward to derive because we can write the joint density \( f(y_{2i}, y_{1i}|x_i) \) as \( f(y_{1i}|y_{2i}, x_i) f(y_{2i}|x_i) \). We have that

\[ \ln f(y_{2i}|x_i) = -\frac{1}{2} \left( p \ln 2\pi + \ln |\Sigma_{22}| + v_i' \Sigma_{22}^{-1} v_i \right) \]

and

\[ \ln f(y_{1i}|y_{2i}, x_i) = \begin{cases} 
 \ln \left\{ 1 - \Phi \left( \frac{m_i - a}{\sigma_{u|v}} \right) \right\} & y_{1i} = a \\
 -\frac{1}{2} \left\{ \ln 2\pi + \ln \sigma^2_{u|v} + \frac{(y_{1i} - m_i)^2}{\sigma_{u|v}^2} \right\} & a < y_{1i} < b \\
 \ln \Phi \left( \frac{m_i - b}{\sigma_{u|v}} \right) & y_{1i} = b 
\end{cases} \]

where

\[ m_i = z_i \delta + (y_{2i} - x_i' \Pi) \Sigma_{22}^{-1} \Sigma_{12} \]

and \( \Phi(\cdot) \) is the normal distribution function so that the log likelihood for observation \( i \) is

\[ \ln L_i = w_i \{ \ln f(y_{1i}|y_{2i}, x_i) + \ln f(y_{2i}|x_i) \} \]

where \( w_i \) is the weight for observation \( i \) or one if no weights were specified. Instead of estimating \( \sigma_{u|v}^2 \) and \( \sigma_v^2 \) directly, we estimate \( \ln \sigma^2_{u|v} \) and \( \ln \sigma_v^2 \).

With maximum likelihood estimation, this command supports the Huber/White/sandwich estimator of the variance and its clustered version using \texttt{vce(robust)} and \texttt{vce(cluster clustvar)}, respectively. See \texttt{[P] _robust}, particularly \texttt{Maximum likelihood estimators} and \texttt{Methods and formulas}.

The maximum likelihood version of \texttt{ivtobit} also supports estimation with survey data. For details on VCEs with survey data, see \texttt{[SVY] Variance estimation}.

The two-step estimates are obtained using Newey’s (1987) minimum chi-squared estimator. For more details on the minimum chi-squared estimator, see \texttt{[R] ivprobit}.
Acknowledgments

The two-step estimator is based on the tobitiv command written by Jonah Gelbach of the University of Pennsylvania Law School and the ivtobit command written by Joe Harkness formerly with the Institute of Policy Studies at Johns Hopkins University.

References


Also see

[R] ivtobit postestimation — Postestimation tools for ivtobit

[R] gmm — Generalized method of moments estimation

[R] ivprobit — Probit model with continuous endogenous covariates

[R] ivregress — Single-equation instrumental-variables regression

[R] regress — Linear regression

[R] tobit — Tobit regression

[ERM] eintreg — Extended interval regression

[SVY] svy estimation — Estimation commands for survey data

[XT] xtintreg — Random-effects interval-data regression models

[XT] xttobit — Random-effects tobit models

[U] 20 Estimation and postestimation commands